BBF RFC 105: The Intein standard - a universal way to modify proteins after translation

Constantin Ahlmann-Eltze, Charlotte Bunne, Magdalena Büscher, Jan Gleixner, Max Horn, Anna Huhn, Nils Klughammer, Jakob Kreft, Elisabeth Schäfer, Carolin Schmelas, Silvan Schmitz, Max Waldhauer, Philipp Bayer, Stephen Krämer, Julia Neugebauer, Pierre Wehler, Joel Beaudouin, Barbara Di Ventura, Roland Eils

October 26, 2014

1 Purpose

This Request for Comments (RFC) proposes a new standard that allows for easy and flexible cloning of intein constructs and thus makes this technology accessible to the synthetic biology community.

2 Relation to other BBF RFCs

RFC[105] does not replace any earlier BBF RFC completely. As RFC[105] describes a way for the assembly of fusion proteins containing one or two intein parts, it replaces fusion protein cloning standards such as BBF RFCs 12[1], 23[2], 25[3], 26[4] and 37[5] for that special case. RFC[105] is fully compatible with and thus extends RFC[10][6] or any other cloning standard not facilitating BsaI.

3 Copyright Notice

4 Nomenclature and Abbreviations

RFC refers to a BioBrick Foundation (BBF) Request for Comments (RFC). All sequences herein are denoted 5' to 3'. A Part – written with capital P – represents a piece of DNA being a functional unit, as meant by the Registry of Biological Parts[7]. X^N denotes the N-terminal part of the protein or fusion

protein X. CDS refers to a coding sequence without stop codons. POI refers to a protein/peptide of interest.

5 Motivation

Inteins are an amazing tool for synthetic biology: their ability for autocatalytic modification of 1-D protein structures allows their use for a great variety of applications ranging from the purification of proteins, protein circularization, protein labeling and biosensing to the direct control of protein activity.[8]

Each of these applications requires the combination of a protein/peptide of interest with a set of inteins or split inteins, which have been characterized previously. Standardization of the cloning process would enable the generation of such constructs from a very limited set of progenitor Parts. Modularization would allow the reuse of existing Parts and the easy exploration of different inteins or exteins for a given task.

However, existing standards do not meet the requirements for such a standard, as they either leave scars at the splice site (RFCs[12[1], 23[2], 25[3], 37[5]]), which compromise the functionality of the final constructs, or require custom oligos (RFC[26[4]]) and tedious backbone amplification(RFCs[28[9], 53[10], 61[11]]) for each intein/extein combination.

Therefore, a new standard that allows for the easy and flexible cloning of intein Parts based on six standardized overhangs and type IIs restriction enzymes was developed, tested and will be described in the following.

6 Formal Description

RFC[105] describes the following nine types of constructs i.e. RFC[10] Parts, each consisting of combinations of six sub-parts connected with one of eight standard overhangs.

6.1 Parts

A **Circularization/Oligomerization Part** MUST contain a **C-intein sub-Part** immediately followed by an **insertion site**, immediately followed by an **N-intein sub-Part**. The **C-intein sub-Part** MUST be in-frame with an ORF of the preceding elements. The preceding

elements SHOULD be either a RFC[10] prefix followed by an RBS and a start codon or a RFC[10] prefix for CDSs followed by an `ATG` start codon. The **N-intein sub-Part** SHOULD be followed by two `TAA` stop-codons and the RFC[10] suffix. The sequence MUST not contain any BsaI recognition sites other than the ones specified. See fig. 1 for a graphical explanation.

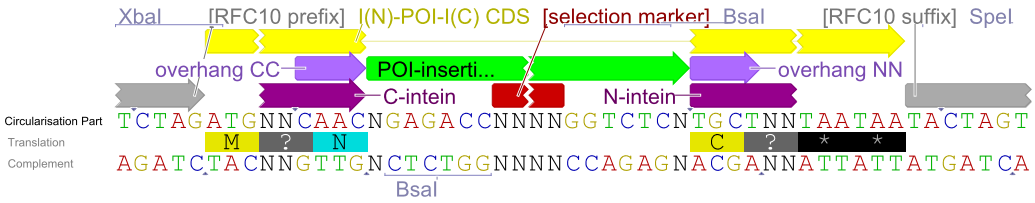


Figure 1: **Circularization/Oligomerization Part**. Optional parts are annotated with squared brackets. (visualized using [12])

An **insert for Circularization/Oligomerization** MUST contain the CDS of a POI preceded by a BsaI recognition site, any single nucleotide and **overhang CC** (`GGTCTCNCAAC`) and followed by **overhang NN**, any single nucleotide and the reversed BsaI recognition site (`TGCTNGAGACC`). It MAY be surrounded by RFC[10] prefix and suffix.

A **N-intein assembly Part** MUST contain the nucleotides `GATG` (**overhang A**) immediately followed by an **insertion site**, immediately followed by an **N-intein sub-Part**. The **N-intein sub-Part** should be followed by two `TAA` stop-codons and the RFC[10] suffix. A His6 tag or any other additional CDS MAY be put before the stop codons. The **overhang A** should be preceded by either the RFC[10] prefix and a RBS or the RFC[10] prefix for CDSs shortened by the final `G`. The sequence MUST not contain any BsaI recognition sites other than the ones specified. See fig. 2 for a graphical explanation.

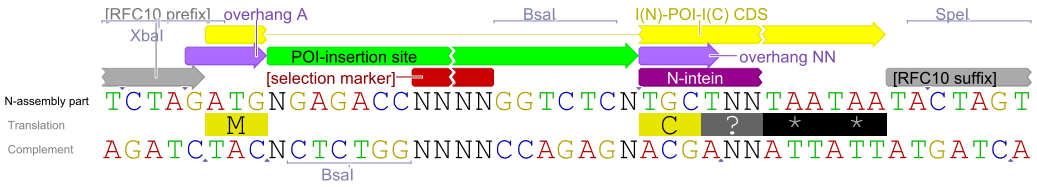


Figure 2: **N-intein assembly Part** Optional parts are annotated with squared brackets. (visualized using [12])

An **N-intein insert** MUST contain the CDS of a POI preceded by a BsaI recognition site, any single nucleotide and **overhang A** (`GGTCTCNGATG`) and followed by **overhang NN**, any single nucleotide and the reversed BsaI recognition site (`TGCTNGAGACC`). The sequence SHOULD not contain any BsaI recognition sites other than the ones specified. It MAY be surrounded by RFC[10] prefix and suffix.

A **N-intein assembly Part with additional C-terminal insertion site** is a **N-intein assembly Part** that MUST have an additional insertion site preceded by `NNTGGT` (**overhang NC**) between the **N-intein sub-Part** and the two `TAA` stop codons. The arbitrary `NN` nucleotides SHOULD be `GG` so that the six additional nucleotides code for two Gly.

A **C-intein assembly Part** MUST contain a **C-intein sub-Part**, immediately followed by an **insertion site**, immediately followed by `TAAT` (**overhang Z**) immediately followed by `AA` resulting in two `TAA` stop-codons. The **C-intein sub-Part** MUST be in-frame with an ORF of the preceding elements. The preceding elements SHOULD be either a RFC[10] prefix followed by an RBS and a start codon or a RFC[10] prefix for CDSs followed by an `ATG` start codon. The sequence MUST not contain any BsaI recognition sites other than the ones specified. See fig. 3 for a graphical explanation.

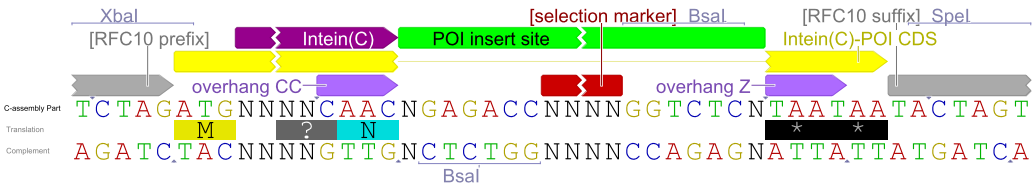


Figure 3: **C-intein assembly Part** Optional parts are annotated with squared brackets. (visualized using [12])

A **C-intein insert** MUST contain the CDS of a POI preceded by a BsaI recognition site, any single nucleotide and **overhang CC** (`GGTCTCNCAAC`) and followed by **overhang Z**, any single nucleotide and the reversed BsaI recognition site (`TAATNGAGACC`). The sequence SHOULD not contain any BsaI recognition sites other than the ones specified. It MAY be surrounded by RFC[10] prefix and suffix.

A **C-intein assembly Part with additional N-terminal insertion site** is a **C-intein assembly Part** that MUST have an additional insertion site preceded by `GATG` (**overhang A**) and followed by `GGTG` (**overhang CN**) and two more nucleotides before the **C-intein assembly Part**. The two nucleotides SHOULD together with the final G of **overhang CN** code for a Gly. `GT` MAY be used.

An **intein insert part** contains a BsaI site followed by any single nucleotide, an **N-intein sub-Part** followed by the corresponding **C-intein sub-Part** followed by any single nucleotide and the reverse BsaI recognition site. The sequence MUST not contain any BsaI recognition sites other than the ones specified. It MAY be surrounded by RFC[10] prefix and suffix.

A **POI with insertion site** MUST have **overhang NN** followed by an **insertion site** and **overhang CC** inserted in frame into its CDS. The sequence SHOULD not contain any BsaI recognition sites other than the ones specified. It SHOULD be a valid RFC[10] **Part**.

6.2 Sub-Parts

An **N-intein sub-Part** is the CDS of an N-intein including the codon for the first amino acid after the splice site, usually a Cys, but not the

codon for the last N-extein amino acid. The first four native nucleotides MUST be changed to `TGCT` (**overhang NN**) (`TGC` coding for Cys). The 5^{th} and 6^{th} nucleotide SHOULD be changed such that they together with 4^{th} nucleotide `T` code for the the native or at least a similar amino acid. For a **non-splicing N-intein sub-Part overhang NN***(`GGCT`) MUST be used instead of **overhang NN**.

A **C-intein sub-Part** is the CDS of a C-intein including the codon for the last amino acid before the splice site, usually an Asn, but not the codon for the first extein amino acid, usually a Cys. The last four native nucleotides MUST be changed to `CAAC` (**overhang CC**) (the `AAC` coding for Asn). The 6^{th} and 5^{th} last nucleotides SHOULD be changed such that they together with 4^{th} last nucleotide `C` code for the the native or at least a similar amino acid. For a **non-splicing C-intein sub-Part overhang NN***(`AGGC`) MUST be used instead of **overhang CC**.

An **insertion site** MUST start with an arbitrary nucleotide followed by the reversed BsaI recognition site (`GAGACC`) and end with an arbitrary nucleotide preceded by the BsaI recognition site (`GGTCTC`). This way BsaI will cut the top strand 4 nucleotides upstream and directly downstream of the insertion site and the bottom strand directly upstream and the bottom strand 4 nucleotides downstream of the insertion site. The insertion site SHOULD contain a **selection marker**. An additional insertion site MAY use a reverse BsmBI (`GAGACG`) and a BsmBI recognition site (`CGTCTC`) instead of the reverse BsaI and the BsaI site respectively.

A **selection marker** MUST be a DNA sequence that if transformed into a cell allows for selection of clones not carrying it. BBa_J04450 is RECOMMENDED.

6.3 Standard overhangs

RFC[105] defines the following eight 4 nt standard overhangs that will allow for flexible assembly of intein fusion proteins from progenitor parts without interfering scars:

A: Sequence: `GATG`. This overhang contains an ATG start codon. It serves as the connection between backbones and N-terminal POIs.

NN: Sequence: `TGCT`. This overhang codes for the first amino acid of N-inteins, a Cys, and for a third of the +2 amino acid, either a Phe, Leu, Ser, Tyr, Cys or Trp.

NN*: Sequence: `GGCT`. Standard overhang for the assembly of non-splicing control Parts. The `GGC` codes for a Gly instead of a Cys which acts as the major nucleophile in the splicing process.

NC: Sequence: `TGGT`. This overhang codes together with the RECOMMENDED preceding `GG` for two Gly. It allows for the insertion of an additional POI behind the C-terminus of N-Inteins.

CC: Sequence: `CAAC`. This overhang codes for an Asn, the last amino acid of C-inteins, and for one third of -2 amino acid. It allows to connect POIs with C-intein CDSs.

CC*: Sequence: `AGGC`. Standard overhang for the assembly of non-splicing control Parts. The `GGC` codes for a Gly instead of Asn which seems to be required for C-terminal cleavage. Like **overhang CC**it is used to connect POIs with C-intein CDSs.

CN: Sequence: `GGTG`. This overhang codes together with the RECOMMENDED following `GT` for two Gly which serve as linker between the N-terminus of a **C-intein sub-Part** and additional POI CDSs.

Z: Sequence: `TAAT`. This overhang builds together with the RECOMMENDED following `AA` two TAA stop codons and serves as the connection between C-terminal POIs and backbones.

7 Usage

Fusion protein Parts with Inteins can be easily assembled from RFC[105] conforming progenitor Parts using a one-pot Golden Gate[13] assembly reaction. The protocol listed in appendix 10.1 MAY be employed to perform this reaction.

If the progenitor parts include additional BsaI sites, the user SHOULD perform an additional religation step after the Golgen Gate assembly reaction. The protocol listed in appendix 10.2 MAY be employed to perform this reaction. In that case, the progenitor part MUST contain a **selection marker**, as the original backbone will recircularize.

Progenitor Parts MAY be assembled using Circular Polymerase Extension Cloning (CPEC)[14] or High Throughput CPE Cloning and Transformation (HiCT), as described in BBF RFC 99[15]. Inserts, if not available on plasmids, MAY be created using an extension PCR reaction, which adds the flanking BsaI sites as described above. It is then RECOMMENDED to also introduce this part into a standard RFC[10] backbone, so that the flanked insert created by the aforementioned extension PCR is available for future assemblies.

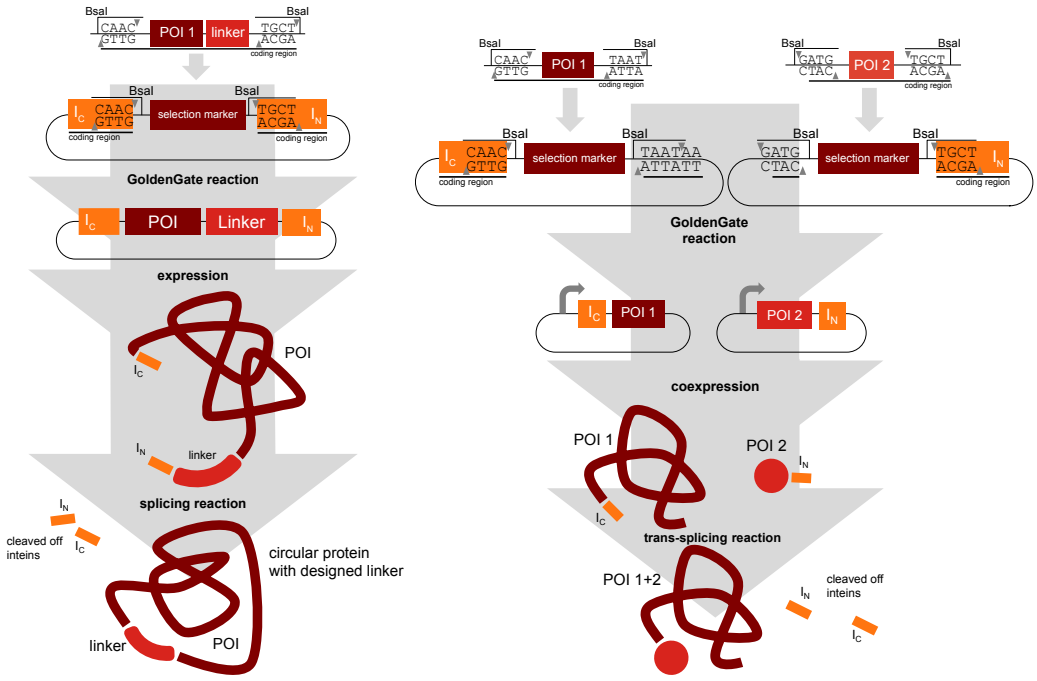
Inteins can be used to circularize or oligomerize proteins by fusing a C-intein to the N-terminus of that protein and the corresponding N-intein to the C-terminus (I^C-POI-I^N). See fig. 4a for a schematic overview of the assembly of such parts using standard constructs. Figure 4b gives an example of how **intein assembly Parts** may be used to add a tag to a POI after translation. A similar strategy can be used for most other intein parts.

8 Discussion

Standardization of overhangs between distinct groups of functionally related parts bears a big advantage. It allows reusing these parts for different assemblies, eliminates the need for the design of custom cloning strategies and eases the overall process. However, they usually come at the cost of interfering scars introduced at the ligation site.

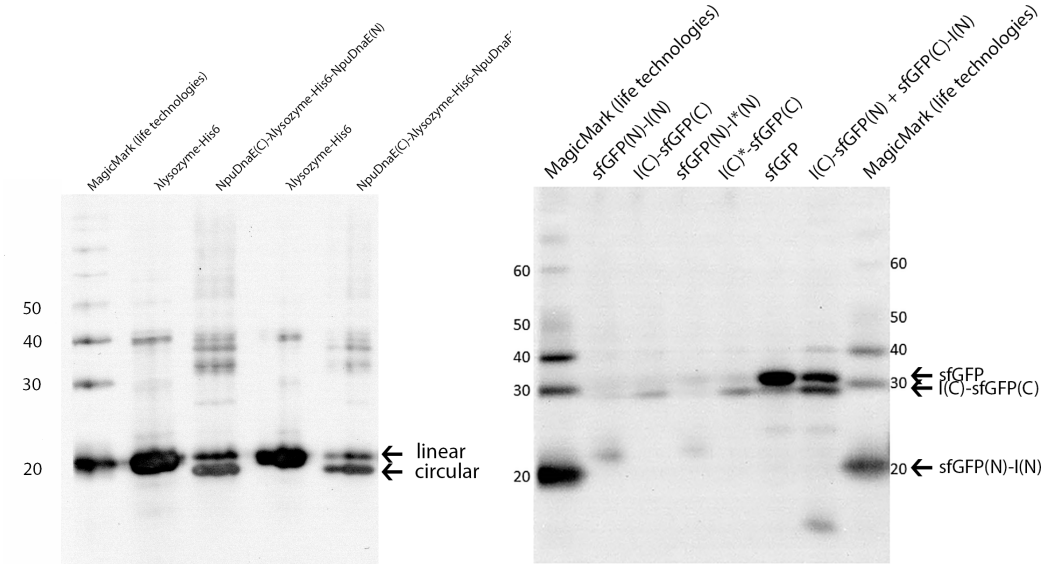
RFC[105]was specifically designed to overcome this problem by defining functional standardized overhangs. The defined overhangs are either part of existing standards (**overhang A**, **overhang Z**), code for short Gly linkers that minimize interference of Tags or other POIs with intein domains (**overhang NC**, **overhang CN**) or are highly conserved motifs in intein sequences (**overhang NN**, **overhang CC**). The latter are most relevant for the intein function: Additional amino acids that would be introduced by a scar would also appear in the spliced proteins and thus hinder the design of optimal linkers for circularization and render protein activation by reconstitution of the active site of enzyme impossible.

Overhang NN can code for the N-terminus of 50.3 % of all N-inteins listed in the intein database InBase[16]. This does not decrease the potential of the proposed method, since those amino acids could be substituted by chemically similar ones and the most promising inteins like Npu DnaE or gp41-1 are compatible. The aspartic acid at the C-terminus of C-inteins is even more conserved so that 93.1% of all intein C-termini resemble **overhang**



(a) Protein circularization (b) Protein assembly

Figure 4: Usage example of standard circularization and assembly Parts. (a) Use of the **Circularization/Oligomerization Part** with a **insert for Circularization/Oligomerization** for the generation a circular protein. (b) Usage of the **N-intein assembly Part** with an **N-intein insert** and the **C-intein assembly Part** with a **C-intein insert**.



(a) Anti-His Western blot image that shows the circularization of λ-lysozyme using an RFC[105] Npu DnaE circularization Part.

(b) Western blot image that shows the reconstitution of split sfGFP using the Npu DnaE split intein.

Figure 5: Example results of applications of standard circularization and assembly Parts.

CC.

The described standard was extensively used by the iGEM Team Heidelberg for the circularization of several enzymes and the reconstitution of split fluorescence proteins. Moreover an intein toolbox was build around it. For instance BBa_K1362000, BBa_K1362100 and BBa_K1362101 represent respectively a RFC[105] conform **Circularization/Oligomerization Part**, **N-intein assembly Part** and **C-intein assembly Part** utilizing the Npu DnaE intein. Figure 5 shows exemplary results of the successful application of the herein described methods. Visit the team wiki (`http://2014.igem.org/Team:Heidelberg`) for detailed information.

9 Authors' Contact Information

Constantin Ahlmann-Eltze: `constantin.ahlmann@t-online.de`
Charlotte Bunne: `bunne@stud.uni-heidelberg.de`

Magdalena Büscher: m.buescher@dkfz-heidelberg.de
Jan Gleixner: jan.gleixner@gmail.com
Max Horn: maexlich@gmail.com
Anna Huhn: anna.g.huhn@gmail.com
Nils Klughammer: klughammer@stud.uni-heidelberg.de
Jakob Kreft: jakob@kreft-mail.de
Elisabeth Schäfer: lisl@schaeferhome.de
Carolin Schmelas: schmelas@stud.uni-heidelberg.de
Silvan Schmitz: silvan@silvanschmitz.de
Max Waldhauer: waldhauer@stud.uni-heidelberg.de

Philipp Bayer: philipp.bayer@lab-alumni.de
Stephen Krämer: stephenkraemer@gmail.com
Julia Neugebauer: julia.neugebauer@bioquant.uni-heidelberg.de
Pierre Wehler: pierre.wehler@bioquant.uni-heidelberg.de

Joel Beaudouin: j.beaudouin@dkfz.de
Barbara Di Ventura: barbara.diventura@bioquant.uni-heidelberg.de
Roland Eils: r.eils@dkfz.de

References

1. Knight, T. BBF RFC-12 Draft Standard for Biobrick BB-2 Biological Parts. *BioBricks Found. Req. Comments.* <http://hdl.handle.net/1721.1/45138> (2007).
2. Phillips, I. & Silver, P. A New Biobrick Assembly Strategy Designed for Facile Protein Engineering. *BioBricks Found. Req. Comments.* <http://hdl.handle.net/1721.1/32535> (2006).
3. Müller, K., Arndt, K., iGEM 2007 Team Freiburg & Grünberg, R. BBF RFC 25: Fusion Protein (Freiburg) Biobrick assembly standard. *BioBricks Found. Req. Comments.* <http://hdl.handle.net/1721.1/45140> (2009).
4. Sleight, S. C. BBF RFC 26: In-Fusion BioBrick Assembly. *BioBricks Found. Req. Comments.* <http://hdl.handle.net/1721.1/46328> (2009).

5. Benčina, M. & Jerala, R. BBF RFC 37 : Fusion protein BioBrick assembly standard with optional linker extension. *BioBricks Found. Req. Comments.* doi:http://hdl.handle.net/1721.1/73912. <http://hdl.handle.net/1721.1/46705> (2009).

6. Knight, T. *Draft Standard for Biobrick Biological Parts* 2007. <http://hdl.handle.net/1721.1/45138>.

7. Selvarajah, V. *What are biological parts* 2013.

8. Wood, D. W. & Camarero, J. a. Intein applications: from protein purification and labeling to metabolic control methods. *J. Biol. Chem.* **289,** 14512–9. ISSN: 1083-351X (May 2014).

9. Peisajovich, S. G., Horwitz, A., Hoeller, O., Rhau, B. & Lim, W. BBF RFC 28: A method for combinatorial multi-part assembly based on the Type IIs restriction enzyme AarI. *BioBricks Found. Req. Comments,* 1–5 (2009).

10. Jiang, H. *et al.* BBF RFC 53: USTC MetaPart Assembly Standard – Extending RFC 10 to Enable Scarless Protein Fusion with Type IIS Restriction Enzyme EarI and SapI. *BioBricks Found. Req. Comments,* 1–28 (2010).

11. Shi, Z., Li, T. & Chen, G. BBF RFC 61: Fast multiple gene fragment ligation method based on Type IIs restriction enzyme DraIII. *BioBricks Found. Req. Comments.* <http://hdl.handle.net/1721.1/59802> (2010).

12. Biomatters. *Geneious version 7.1* <http://www.geneious.com>.

13. Engler, C., Kandzia, R. & Marillonnet, S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE* **3,** e3647 (Nov. 2008).

14. Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PloS one* **4,** e6441 (2009).

15. Beer, R. *et al.* BBF RFC 99: HiCT: High Throughput Protocols For CPE Cloning And Transformation (2013).

16. Perler, F. B. InBase: the Intein Database. *Nucleic Acids Res.* **30,** 383–4. ISSN: 1362-4962 (Jan. 2002).

10 Appendix: Recommended materials and methods

10.1 Golden Gate Assembly (cycling)

- Mix 150 ng of the backbone and equimolar amounts of the insert(s) in water in a PCR tube for a total volume of 15 µL.
- Add 1.5 µL of 10X T4 Ligase Buffer and, when using BsaI or another restriction enzyme that requires it, BSA (Bovine Serum Albumin) at a final concentration of 1X.
- Add 1 µL of each restriction enzyme and 1 µL of T4 DNA Ligase (400,000 cohesive end ligation units/ml).
- Optional: Add 1 µL of T4 Polynucleotide Kinase, if several inserts without 5'-phosphorylation (e.g. annealed oligos) are used.
- Place the reaction in a thermocycler and run the following program: 25 cycles of 4 min ligation at 16 °C, 3 min restriction at 37 °C, 5 min at 50 °C (final restriction) and 5 min of heat inactivation at 80 °C.

10.2 Religation after Golden Gate Assembly

- Add 12.5 µL of water, 1.5 µL of T4 Ligase Buffer and 1 µL of T4 DNA Ligase to your Golden Gate reaction.
- Incubate for 20 min at 16 °C, then heat inactivate for 10 min at 65 °C.